Model Transport: Towards Scalable Transfer Learning on Manifolds

Freifeld, Oren ; Hauberg, Søren; Black, Michael J.

Published in:
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014

Link to article, DOI:
10.1109/CVPR.2014.179

Publication date:
2014

Document Version
Peer reviewed version

Model Transport: Towards Scalable Transfer Learning on Manifolds

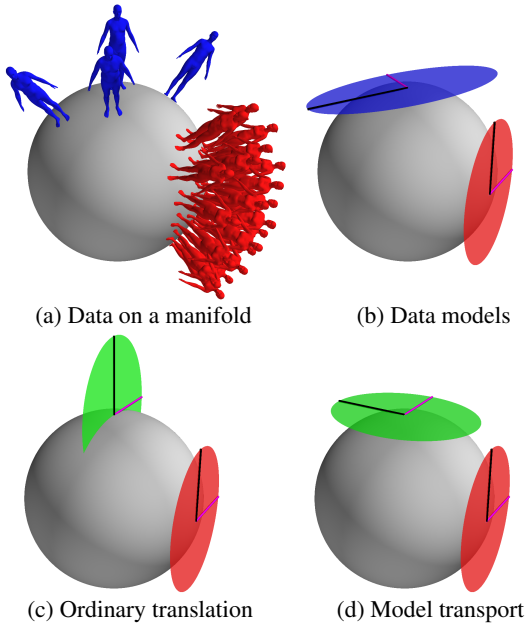
Oren Freifeld
MIT
Cambridge, MA, USA
freifeld@csail.mit.edu

Søren Hauberg
DTU Compute
Lyngby, Denmark
sohau@dtu.dk

Michael J. Black
MPI for Intelligent Systems
Tübingen, Germany
black@tue.mpg.de

Abstract

We consider the intersection of two research fields: transfer learning *and* statistics on manifolds. *In particular, we consider, for manifold-valued data, transfer learning of tangent-space models such as Gaussians distributions, PCA, regression, or classifiers. Though one would hope to simply use ordinary $\mathbb{R}^n$-transfer learning ideas, the manifold structure prevents it. We overcome this by basing our method on inner-product-preserving* parallel transport, *a well-known tool widely used in other problems of statistics on manifolds in computer vision. At first, this straightforward idea seems to suffer from an obvious shortcoming: Transporting large datasets is prohibitively expensive, hindering scalability. Fortunately, with our approach,* we never transport data. *Rather, we show how the* statistical models *themselves can be transported, and prove that for the tangent-space models above, the transport "commutes" with learning. Consequently, our compact framework, applicable to a large class of manifolds, is not restricted by the size of either the training or test sets. We demonstrate the approach by transferring PCA and logistic-regression models of real-world data involving 3D shapes and image descriptors.*



(a) Data on a manifold (b) Data models (c) Ordinary translation (d) Model transport

Figure 1: Model Transport for covariance estimation. On nonlinear manifolds, statistics of one class (red) are transported to improve a statistical model of another class (blue). While ordinary translation is undefined (c), and data transport is expensive, a model can be inexpensively transported (green) while preserving data statistics (d).

1. Introduction

In computer vision, manifold-valued data arise often. The advantages of representing such data explicitly on a manifold include a *compact* encoding of constraints, *distance* measures that are usually superior to ones from $\mathbb{R}^n$, and *consistency*. For such data, statistical modeling on the manifold is generally better than statistical modeling in a Euclidean space [12, 18, 29, 39]. Here we consider the first scalable generalization, from $\mathbb{R}^n$ to Riemannian manifolds, of certain types of transfer learning (TL). In particular, we consider TL in the context of several popular *tangent-space models* such as Gaussian distributions (Fig. 1b), PCA, classifiers, and simple linear regression. In so doing, we recast TL on manifolds as TL between tangent spaces. This generalizes those $\mathbb{R}^n$-TL tasks where models learned in one region of $\mathbb{R}^n$ are utilized in another. Note, however, that we do not claim that all $\mathbb{R}^n$-TL tasks have this form.

Let M denote an n-dimensional manifold and let T_pM and T_qM denote two tangent spaces to M, at points $p, q \in M$. One cannot simply apply models learned in T_pM to data in T_qM as these, despite both being isomorphic to $\mathbb{R}^n$, are two different spaces: A model on T_pM is usually not even defined in T_qM; see Fig. 1c. Such obstacles, caused by the curvature of M, do not arise in $\mathbb{R}^n$-TL. To address this we could parallel transport (PT) [7] the data from T_pM to T_qM, learn a model of the transported data in T_qM, and use

this model in T_qM. In fact, this brings us back to the setting of ordinary $\mathbb{R}^n$-TL whose methods are now transparently applicable. The question whether this idea is statistically useful depends on the application and the data of interest; our experiments show that the answer is often positive.

Unfortunately, this solution scales poorly: Transporting *data* is expensive. This is particularly true for large datasets or when the PT has no closed form. Also, the value of q (which determines T_qM) is sometimes only available at test time, and neither the transport nor the learning can be done offline. Our key contribution is to prove that, for a fairly large class of important models, *one never has to transport the data*. Rather, not only is it possible to *transport the model* but we also prove that the resulting model is identical to what would have been gained from the expensive data-transport-based approach mentioned above. Loosely speaking, we prove that *learning commutes with parallel transport*. Consequently, the computational complexity of our framework is not proportional to the size of the *dataset*; rather, it is proportional to the complexity of the *model*. It also frees us from having to store or access the training data during test time.

Our *model transport* (MT) framework is straightforward to apply yet it allows us to solve seemingly difficult problems with surprising results; *e.g.*, we find that on a manifold of mesh deformations [12], observations of people with normal Body-Mass Index (BMI) help modeling high-BMI people. Similarly, we show that the transported shape model of women, learned from many examples, improves a shape model of men learned from only few examples. We also show that MT is useful for classifiers of image descriptors encoded as Symmetric Positive-Definite (SPD) matrices [40]. To conclude, we present a framework that will enable researchers to apply $\mathbb{R}^n$-TL ideas to manifold-valued data while being transparently-scalable to large datasets.

2. Previous Work

Manifold-valued data are ubiquitous in computer vision. The $SE(3)$ and $SO(3)$ groups are omnipresent while spherical data appear in omnidirectional images [26], normalized features [20], pre-shape spaces [21] and surface normals [37]. Articulated poses are represented by rotations or length constraints [18, 27]. A square root of a probability density function is a point on the unit sphere [35]. Anatomical variability is captured via certain Lie groups [15]. Diffusion-tensor images yield orientation distribution functions (*e.g.* [8]) or SPD matrices (*e.g.*, [1]). The latter are also used as image descriptors [40]. Flow patterns are described by matrix Lie groups [23, 42] while infinite-dimensional diffeomorphisms are used for nonrigid image deformations (*e.g.* [2]). Kendall's shape spaces are also manifolds [21]. The Grassmannian captures affine shapes [3, 34, 39]. Mesh deformations enable a tractable generative framework for deformable triangular surfaces [12]. See [36, 44] for additional shape spaces.

Statistics on manifolds is often done via parametric models on tangent spaces, *e.g.* [11, 12, 17, 28, 33, 37, 39, 41], but alternatives (not considered here) exist, *e.g.* [4, 10, 19, 38]. *Statistics on manifolds* (our work included) is different from *manifold learning*, where a low-dimensional latent manifold is learned from data in $\mathbb{R}^n$. In contrast, here M is known and the goal is to model statistics of M-valued data.

We do not advocate a specific manifold or a specific Riemannian metric and we do not seek to define a new tangent-space model. Rather, we strive for an efficient TL framework that applies to many manifolds and many models. We base our method on PT but emphasize we are not the first to use PT in statistical computer-vision tasks: PT has been used in other, non-TL, applications; *e.g.* [24, 25, 30, 39]. Our approach differs from such work in, not only the application, but also how PT is used: We PT models, not data.

Transfer Learning: Using $\mathbb{R}^n$-valued data from one dataset in an inferential task on another is a classical TL problem (*e.g.* [5]). A model learned from one set can be a prior for a model learned from the second [22]. Models can also be learned independently and then be combined [32]. One may also use two mixture models, one per set, which share a component [6]. There are many other examples of TL; a full review is beyond our scope. Surprisingly, despite the success of TL in $\mathbb{R}^n$, the ubiquity of manifold-valued data, and the effectiveness of statistics on manifolds, $\mathbb{R}^n$-TL solutions have yet to be generalized to manifolds. Note that our M-valued data setting should not be confused with that of [14], where, for $\mathbb{R}^n$-valued data, they exploit the geometry of a space of models. Also, in *manifold alignment* [16], TL is done via estimating a latent low-dimensional manifold shared by two $\mathbb{R}^n$-valued datasets, while in our setting the data lie on a known manifold, and our goal is TL across it. Wei *et al.* [42] share information across a Lie group using PT. Particularly, they PT bases that span linear subspaces of the Lie algebra. Unfortunately, their choice of PT does not preserve inner-products so second-order statistics are distorted. Thus, they cannot transport covariances, and their transported PCA models are distorted. Our approach does not have this problem, is applicable to more models (*e.g.*, regression and classification) and does not require M to be a Lie group. Hauberg *et al.* [17] transport covariances in a tracking application. Our method differs in the application, applies to a broader class of models, and also sheds new light on their procedure, by showing it can be justified in a concrete statistical sense. Closely related to ours, is a very recent work by Xie *et al.* [43] who use PT for PCA models in a shape space (whose elements are shapes, not shape transformations). Their work is not focused on MT in general, and while they apply PT to PCA eigenvectors,

they do not justify why this is optimal. Our work does not only fill this theoretical gap but also treats PT of regressions/classifiers which are not covered in [43].

3. Mathematical Background

Assuming familiarity with basic Riemannian geometry (see [7] for an introduction), we sketch the additional background required for our method. Henceforth, M is a geodesically-complete Riemannian manifold of dimension n, and $d : M \times M \to \mathbb{R}^+$ is a geodesic distance on M. If $p \in M$, we let Exp_p and Log_p denote its associated Riemannian exponential and logarithmic maps.

3.1. Statistics on Manifolds

We review relevant notions in statistics on manifolds [11, 28]. Various concepts from classical statistics in $\mathbb{R}^n$ can be generalized to M. A popular approach utilizes tangent spaces. Let $\{p_i\}_{i=1}^N \subset M$. The *Fréchet function*, $Q : M \to \mathbb{R}^+$: $p \mapsto \sum_{i=1}^N d(p, p_i)^2$, generalizes the $\mathbb{R}^n$-notion of sum-of-squared-distances. A local minimizer of Q (called the *Karcher mean*) can usually be efficiently computed [28]. Let μ denote the Karcher mean. A covariance is defined via the covariance of the data as expressed in $T_\mu M$: $\mathrm{Cov}(\{p_i\}_{i=1}^N) \triangleq \frac{1}{N-1}\sum_{i=1}^N \mathrm{Log}_\mu(p_i)\mathrm{Log}_\mu(p_i)^T$. As $T_\mu M \cong \mathbb{R}^n$, PCA can be done on $\{\mathrm{Log}_\mu(p_i)\}_{i=1}^N$ [11]. If the p_i's are labeled, and $\{\mathrm{Log}_\mu(p_i)\}_{i=1}^N$ are seen as the independent variable, then we can define a simple linear regression (for $\mathbb{R}$-valued labels) or a logistic regression (for binary labels)[1]. Our treatment of both cases is similar and we may assume, without loss of generality, that the labels are real. Such problems should not be confused with those where labels are M-valued, but the independent variable is real (*e.g.* [10]).

3.2. Parallel Transport

In $\mathbb{R}^n$, data and models can be moved from one region to another via ordinary translation. On M, this fails; *e.g.* if $x \in T_pM$ then $x+q-p$ is rarely in T_qM. Similarly, models cannot be naively translated; *e.g.* a covariance, which can be viewed as a bilinear form over one tangent space, cannot be simply translated to another (Fig. 1c). The same applies for linear regression (a linear functional over a particular tangent space) and other models. However, vectors can be moved from T_pM to T_qM using a well-known tool called *parallel transport* (PT), to be defined as follows[2]. Suppose every smooth curve $c : [0, 1] \to M$ is associated with a collection of maps, $\left\{\Gamma^c_{s,t} : T_{c(s)}M \to T_{c(t)}M | s, t \in [0, 1]\right\}$, such that: (1) $\Gamma^c_{s,s}$ is the identity map on $T_{c(s)}M$; (2) $\Gamma^c_{u,t} \circ \Gamma^c_{s,u} = \Gamma^c_{s,t}$; (3) $\Gamma^c_{s,t}$ depends smoothly on s and t. In which case, if $x \in T_{c(s)}M$, we call $\Gamma^c_{s,t}(x)$ the PT of x along c to $T_{c(t)}M$. PT provides a principled way to move data (when expressed as tangent vectors) across M [24, 25, 39]. In what follows, c, s, and t are either clear from the context or their particular values are immaterial to the discussion, and thus we use Γ as short for $\Gamma^c_{s,t}$; when it is also clear which Γ is used, we write $\widetilde{x}$ instead of $\Gamma(x)$.

Criteria (1-3) are met by various collections of maps and so there exist *many types of PT*. A *metric parallel transport* (MPT) is one that preserves inner-products: $\langle x, y\rangle_p = \langle \widetilde{x}, \widetilde{y}\rangle_q$ for every $x, y \in T_pM$. Thus, orthogonality and distance between vectors are preserved. This criterion too can be met in various ways and so there exist various types of MPT. Note that an MPT preserves second-order statistics: when applied to T_pM-valued random variables, their variances and correlations are unchanged. Also, every MPT is an invertible linear map; *cf.* [13]. Finally, given the MPT of choice, there still remains the technical issue of how to compute it. The solution may be analytical or only numerical, depending on the case. See also discussion in [24, 25].

4. Parallel Transport for Transfer Learning

4.1. The Euclidean Setting

Before the manifold setting, we start with a simpler Euclidean one. Consider the following typical TL tasks.

Task I: The training set consists of two $\mathbb{R}^n$-valued datasets: $\{x_i^{\mathrm{L}}\}_{i=1}^{N_{\mathrm{L}}}$; $\{x_j^{\mathrm{S}}\}_{j=1}^{N_{\mathrm{S}}}$; $N_{\mathrm{L}} > N_{\mathrm{S}}$. Let $p(x^{\mathrm{L}})$ and $p(x^{\mathrm{S}})$ be the unknown generating distributions. Our interest is in modeling $\{x_j^{\mathrm{S}}\}$ using either a normal distribution or PCA. Suppose N_{S} is sufficiently large to estimate the mean of $p(x^{\mathrm{S}})$, but that it is too small to yield a reliable estimate of a covariance or a PCA subspace.

Task II: The training set consists of two $\mathbb{R}^n$-valued datasets, $\{x_i^A\}_{i=1}^{N_A}$ and $\{x_j^B\}_{j=1}^{N_B}$, the x_i's being labeled; N_B need not be smaller than N_A. With $y_i^A = \mathit{label}(x_i^A)$, the y_i^A's are either binary (more generally, categorical) or real[3]. The task is either classification or regression; *i.e.*, to predict $\{y_j^B\}_1^{N_B}$, the labels of x_j^B. We let $p(y^A|x^A)$ and $p(y^B|x^B)$ denote the unknown conditional distributions. A possible approach is to employ a logistic or linear regression; as the mathematical treatment is similar we focus on the latter.

Solutions: In both tasks, we wish to leverage statistics of one dataset when solving an inferential task for another. If $p(x^{\mathrm{L}})$ and $p(x^{\mathrm{S}})$ (respectively, $p(y^A|x^A)$ and $p(y^B|x^B)$) are unrelated, then this would fail. Suppose, however, that there does exist an unknown relation; *e.g.*, $p(x^{\mathrm{L}})$ and $p(x^{\mathrm{S}})$ may be well-modeled by two Gaussians of different means, but with similar covariances. Likewise, the optimal linear-regression models implied by $p(y^A|x^A)$ and $p(y^B|x^B)$ may differ mostly in their intercepts while the associated hyperplanes are similar[4]. Thus, in Task I, following an appropri-

[1] Additional models, *e.g.* SVM [39], can be defined similarly.

[2] Another way to define it is via a *connection* [7].

[3] A close variant is if we have some y_j^B's, but their number is smaller.

[4] In both cases, by "similar" we do not necessarily mean "identical".

ate offset implied by the means, we may fuse the associated Gaussian or PCA models. In Task II, we may learn a model from the (x_i^A, y_i^A) pairs, and then shift the associated hyperplane to the region of $\mathbb{R}^n$ where the x_j^B's reside (we may also follow this by some adaptation).

4.2. The Manifold Setting

This work considers analogous tasks but in the more general and challenging manifold setting. Specifically, we express each M-valued dataset in the tangent space at its Karcher mean. In Task I, given two M-valued datasets, $\{p_i\}_{i=1}^{N_{\mathrm{L}}}$ and $\{q_j\}_{j=1}^{N_{\mathrm{S}}}$, letting p and q denote the respective means, set $x_i^{\mathrm{L}} = \mathrm{Log}_p(p_i)$ and $x_j^{\mathrm{S}} = \mathrm{Log}_q(q_j)$. In Task II, given $\{p_i\}_{i=1}^{N_A}$ and $\{q_j\}_{j=1}^{N_B}$ in M, set $x_i^A = \mathrm{Log}_p(p_i)$ and $x_j^B = \mathrm{Log}_q(q_j)$, reusing the symbols p and q to denote the means (the labels are still either binary or real). The goals are as before, but now the space of interest is T_qM, while the statistics we hope to leverage are in T_pM. Thus we cannot immediately use the latter in the the former.

4.3. Solutions Can Be Based on Parallel Transport – But What Objects Should We Transport?

Any type of PT can be used to move data from T_pM to T_qM but for TL we restrict the choice to the MPT class in order to preserve second-order statistics. More generally, there are many ways, including non-parallel, to move data from T_pM to T_qM, but all create some distortion due to the curvature of M. Without prior knowledge it is sensible to pick the approach which distorts the data the least. This is provided by a particular choice of MPT, called the *Levi-Civita* (LC) PT [7] which is widely used in computer vision. When prior information is available it may, however, be better to pick another MPT that utilizes it (see [24, 30] for examples in image registration tasks). Our framework holds for all choices of MPT. Fix an MPT Γ, and let C denote the computational cost associated with transporting a single vector. C increases with $\dim(M) = n$ and also depends on M and Γ (particularly, it is higher when Γ is not given in closed form). The solutions below differ in their costs as they apply Γ to different numbers of vectors.

Solution 1 – data transport. The most direct solution is to transport the data from T_pM to T_qM using an MPT. This places the data from T_pM in T_qM while preserving second-order statistics of the data. Unfortunately, the computational cost can be high. If N is the size of the dataset, the cost is NC; *e.g.*, a large N_{L} makes $\Gamma(\{x_i^{\mathrm{L}}\}_1^{N_{\mathrm{L}}})$ expensive. Likewise, using $\Gamma^{-1}(\{x_j^{\mathrm{S}}\}_1^{N_{\mathrm{S}}})$, intending to learn a fused model in T_pM, will not help if a large *test set* in T_qM has to be transported to T_pM in order to apply the model.

Solution 2 – basis transport. If the number of data points N is larger than the manifold dimensionality n we can lower the computational cost by transporting the basis of T_pM. This is done by transporting the natural basis, *i.e.* the columns of the $n \times n$ identity matrix. If L denotes the matrix of the transported basis vectors, then the MPT can be computed by a multiplication by L (*i.e.* a change of basis). This has cost nC as we must transport n basis vectors (plus the cost of the multiplication by an $n \times N$ matrix).

4.3.1 A Third Solution – Model Transport

In practice, both solutions are typically computationally expensive. Fortunately, often there is no need to transport either the data or the basis: it suffices to transport the model. In which case, the number of transported vectors is determined by the model complexity, and is fixed w.r.t. n and N. Below we show how this is done. For Tasks I and II, Propositions 4.1 and 4.2 imply that after being learned in T_pM, a model can be then transported to T_qM and we will get exactly the same results as with the expensive solutions above. We provide the proofs in the supplemental material [13].

Proposition 4.1 (Covariance/PCA Transport)**.** *Let $n' \geq n$ and assume M is embedded in $\mathbb{R}^{n'}$ with a Riemannian metric induced by $\mathbb{R}^{n'}$. Let $p, q \in M$ and let $\{x_i\}_{i=1}^N \subset T_pM$. Let VS^2V^T be the eigen-decomposition of XX^T where $X \triangleq [x_1, \ldots, x_N]$ and let VSU^T be the SVD of X, with $\{S_{i,i}\}_{i=1}^n$, the diagonal entries of S, sorted in a non-increasing order. Let $[v_1, \ldots, v_n]$ denote the columns of V. If $x \in T_pM$, let $\tilde{x} \in T_qM$ denote the transport of x according to some MPT along a smooth curve in M from p to q. Let $\tilde{X} \triangleq [\tilde{x}_1, \ldots, \tilde{x}_N]$ and let $\tilde{V} \triangleq [\tilde{v}_1, \ldots, \tilde{v}_n]$. Then:*

(a) $\tilde{V}SU^T$ is the SVD of $\tilde{X}$. while $\tilde{V}S^2\tilde{V}^T$ is the eigen-decomposition of $\tilde{X}\tilde{X}^T$.

(b) If $k < n$, then the k-dimensional PCA model of $\{\tilde{x}_i\}_{i=1}^N \subset T_qM$ is given by (the eigenvectors) $\{\tilde{v}_i\}_{i=1}^k$ and (eigenvalues) $\{S_{i,i}/\sqrt{N-1}\}_{i=1}^k$.

In words: (a) means that X and $\tilde{X}$ share singular values and right-singular vectors, but the left-singular vectors of $\tilde{X}$ are exactly the parallel-transported left-singular vectors of X. The model in (b) is equivalent to first computing a k-dimensional PCA model (in T_pM) of $\{x_i\}_{i=1}^N$, given by the vectors $\{v_i\}_{i=1}^k$ and standard deviations $\{S_{i,i}/\sqrt{N-1}\}_{i=1}^k$, and then transporting the $\{v_i\}_{i=1}^k$ while keeping the standard deviations unchanged.

Proposition 4.1 substantially extends a previous result [17], derived in the context of the Kalman filter, which states that the parallel transport of a T_pM-covariance yields *some* valid T_qM-covariance. Thus, Proposition 4.1 not only provides the basis for a compact TL framework for transporting covariances and PCA models on manifolds, but also sheds a new light on an existing tracking algorithm; note it also provides further mathematical justification to the method used in [43].

Our next proposition provides a similar result for simple linear regression. But first, we need some preliminaries. Let p and q be in an n-dimensional geodesically-complete Riemannian manifold M. Let the inner-products on T_pM and T_qM (implied by the Riemannian metric) be defined by

$$\langle \cdot, \cdot \rangle_p : T_pM \times T_pM \to \mathbb{R} : (x, y) \mapsto x^T A_p y \quad (1)$$

$$\langle \cdot, \cdot \rangle_q : T_qM \times T_qM \to \mathbb{R} : (x, y) \mapsto x^T A_q y \quad (2)$$

where A_p and A_q are SPD. Let $\{x_i\}_{i=1}^N \subset T_pM$, let $\{y_i\}_{i=1}^N \subset \mathbb{R}$ denote their labels, and let $L : T_pM \mapsto T_qM$ denote the linear transformation associated with some MPT along a smooth curve in M from p to q. A simple linear regression model $T_pM \to \mathbb{R}$ has the following form: $x \mapsto x^T\alpha + \alpha_0 = \langle x, A_p^{-1}\alpha \rangle_p + \alpha_0$. Here $\alpha_0 \in \mathbb{R}$, while α and $A_p^{-1}\alpha$ are regarded[5] as elements of T_pM. Let $l_i : \mathbb{R} \to \mathbb{R}_+$ be a loss function associated with y_i (*e.g.*, $l_i : \hat{y}_i \mapsto (\hat{y}_i - y_i)^2$ is a square loss).

Proposition 4.2 (Simple-Linear-Regression Transport)**.** *Let*

$$(\beta, \beta_0) = \underset{\alpha \in T_pM, \alpha_0 \in \mathbb{R}}{\arg\min} \sum_{i=1}^N l_i(x_i^T\alpha + \alpha_0) \ , \quad (3)$$

and set $\gamma = A_q L A_p^{-1}\beta$ *(note that* $\gamma \in T_qM$*). Then*

$$\gamma = \underset{\delta \in T_qM}{\arg\min} \sum_{i=1}^N l_i((Lx_i)^T\delta + \beta_0) \ . \quad (4)$$

An equivalent version of Proposition 4.2 holds for a logistic-regression model (transported using the same expression we used here for γ); we omit the details. Unlike in Proposition 4.1, in Proposition 4.2 we do not require the presence of an ambient space $\mathbb{R}^{n'}$ nor do we impose a particular metric. In fact, it is easy to prove a slightly different version of Proposition 4.1 where these restrictions are removed. However, showing this requires more notational clutter which might obscure the main idea.

Note that for transporting PCA, the cost is kC (k vectors are transported); for linear/logistic regression, it is C (1 vector). By now it should be apparent that similar results can be derived for many other models. The desired commutativity of learning and transport holds for any model where the data enter the equations only via inner products; *e.g.*, in an SVM model one may transport the support vectors.

4.4. Applying Transported Models

Henceforth we will assume $\Gamma : T_pM \to T_qM$ is done along a geodesic curve. Let $\Sigma_{\mathrm{L}} = \mathrm{cov}(\{p_i\}_i^{N_{\mathrm{L}}})$ and set, by abuse of notation, $\Sigma_\Gamma = \Gamma(\Sigma_{\mathrm{L}})$. We can use Σ_Γ as a covariance on T_qM or, letting $\Sigma_{\mathrm{S}} = \mathrm{cov}(\{q_j\}_j^{N_{\mathrm{S}}})$, we may combine them; *e.g.* using *shrinkage estimation* [32]:

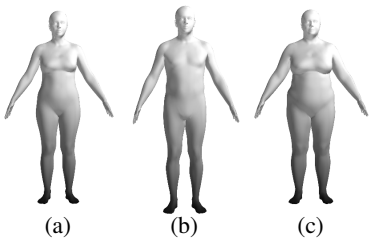
$$\Sigma_\lambda \triangleq \lambda\Sigma_\Gamma + (1-\lambda)\Sigma_{\mathrm{S}} \ , \ 0 \le \lambda \le 1 \ . \quad (5)$$

For PCA, let $V_{\mathrm{L}} \subset T_pM$ and $V_{\mathrm{S}} \subset T_qM$ denote the first k_{L} eigenvectors of Σ_{L} (where $k_{\mathrm{L}} \le N_{\mathrm{L}}$) and first k_{S} eigenvectors of Σ_{S} respectively (we take $k_{\mathrm{S}} = N_{\mathrm{S}}$), and set $V_\Gamma = \Gamma(V_{\mathrm{L}}) \subset T_qM$. Similarly, let $\sigma_{\mathrm{L}}^2 \in \mathbb{R}^{k_{\mathrm{L}}}$ and $\sigma_{\mathrm{S}}^2 \in \mathbb{R}^{k_{\mathrm{S}}}$ denote the eigenvalues. We can now use $(V_\Gamma, \sigma_{\mathrm{L}})$ as a PCA model in T_qM, or combine it with V_{S}; *e.g.*, let $V_{\mathrm{F}} \subset T_qM$ denote an orthonormalized version of $[V_\Gamma, V_{\mathrm{S}}]$. V_{F}, which we regard as a *fused* model, contains $k_{\mathrm{L}} + k_{\mathrm{S}}$ vectors and is able to generalize better than V_{S} as it also contains the transported variation from T_pM. To enable a direct comparison with V_{L} or V_Γ, we can also restrict the combined model to have the same dimensionality by using only k_{L} vectors. Let $X = [\mathrm{Log}_q(q_1), .., \mathrm{Log}_q(q_{N_{\mathrm{S}}})]$, define

$$X_\lambda \triangleq [\lambda V_\Gamma \sigma_{\mathrm{L}}, (1-\lambda)X] \ , \ 0 \le \lambda \le 1 \ , \quad (6)$$

and let V_λ represent the k_{L}-dimensional PCA subspace of X_λ. This is simply weighted PCA, where we treat vectors in V_Γ as examples weighted by the standard deviations. In both Eqs. (5) and (6), the larger λ is, the stronger is the influence of $\{p_i\}_i^{N_{\mathrm{L}}}$. We think of this influence as *regularization*. The value of λ may be chosen by cross-validation.

Transported regression models or classifiers can be utilized in a similar way. More generally, a transported model can either be applied as is in T_qM, or it can be adapted (or fused with a T_qM-model) using $\mathbb{R}^n$-TL methods.



(a) (b) (c)

Figure 2: Mean shapes. These correspond to the Karcher means computed from 1000 female shapes (a), 50 male shapes (b), and 50 shapes of women with high-BMI (c).

5. Results

MT is applicable on many manifolds; here we experiment with two of these, using only real, non-synthetic data.

5.1. PCA Transport and Shape Deformations

For Task I, let M be the manifold of triangular-mesh deformations, proposed in [12]. Points on M are deformations of 3D shapes from a template mesh. M is a Lie group (though it is not a requirement for MT) made out of 21550

[5] Formally, α and $A_p^{-1}\alpha$ are elements of the dual space of T_pM.

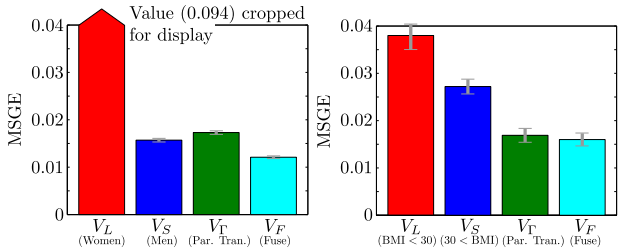
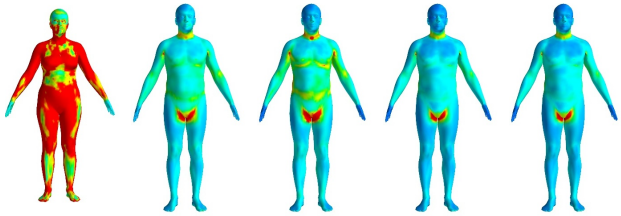


Figure 3: Summary for shape experiments. *Left:* Gender. *Right:* BMI. The bars represent the overall reconstruction error for V_{L}, V_{S}, V_Γ, and V_{F}. For a given model, the height of the bar represents the reconstruction error measured in terms of SGE averaged over the entire test dataset as well as all of the mesh triangles.



(a) V_{L} (b) V_{S} (c) V_Γ (d) V_{F} (e) V_λ

Figure 4: Model mean error: Genders. Blue and red indicate small and large errors respectively. The heat maps are overlaid over the points of tangency associated with the models: p for (a), and q for (b-e). See text for details.

copies of a 6-dimensional Lie group, which is isomorphic to the product of three smaller ones, including $\mathrm{SO}(3)$; thus, $n = 129300$. While here we do not advocate a particular manifold nor does our work focus on shape spaces, this M enables us to easily demonstrate the MT framework. The data consist of aligned[6] 3D scans of real people [31]. On this M, the LC PT is computed as follows: For the $\mathrm{SO}(3)$ components of M, a closed-form solution is available [9], while for the rest we use Schild's ladder (see, *e.g.*, [17, 24]).

From Venus to Mars. We first illustrate the surprising power of MT. The training data contains $N_{\mathrm{L}} = 1000$ shapes of women (Fig. 1a, red; shown here on a 2D manifold for illustration) but only $N_{\mathrm{S}} = 50$ shapes of men (blue), where all shapes are represented as points on M. As it is reasonable to expect some aspect of shape variation among women may apply to men as well, we model the shape variation of men while leveraging that of women. We first compute the Karcher means for women and men denoted p and q, respectively (Fig. 2a–2b). We then compute their PCA models, $V_{\mathrm{L}} \subset T_pM$ and $V_{\mathrm{S}} \subset T_qM$ ($k_{\mathrm{L}} = 200$ and $k_{\mathrm{S}} = 50$), as well as $V_\Gamma = \Gamma(V_{\mathrm{L}})$. For an animated illustration see [13]. We also compute V_{F} and V_λ using the procedures from Sec. 4.4. We evaluate performance on

[6]MT also applies to some shape spaces that do not require alignment.

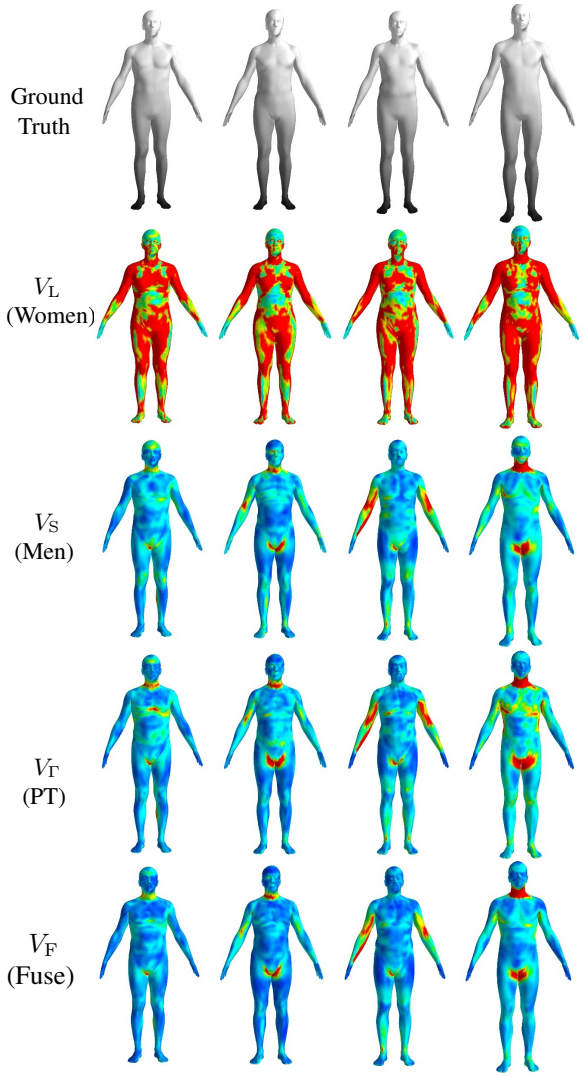
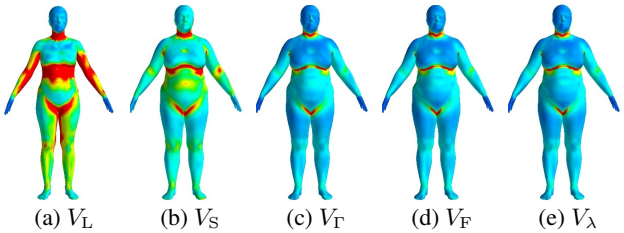


Figure 5: Selected results: Gender. Each column represents a different test body. The heat maps are overlaid on the reconstructions using different models.



(a) V_{L} (b) V_{S} (c) V_Γ (d) V_{F} (e) V_λ

Figure 6: Model mean error: BMI. Analogous to Fig. 4.

1000 test male shapes, whose deformations serve as ground-truth. Let $V \in \{V_{\mathrm{L}}, V_{\mathrm{S}}, V_\Gamma, V_{\mathrm{F}}, V_\lambda\}$. Let μ denote the point of tangency; *i.e.*, p for V_{S} and q otherwise. Let $z_i \in M$ denote the true deformation of test example i. Its reconstruction is $\mathrm{Exp}_\mu(VV^T\mathrm{Log}_\mu(z_i)) \in M$. We then com-

pute, for each triangle, the Squared Geodesic Error (SGE) between the reconstruction and the true deformation. Fixing i, SGE is averaged over all body triangles, yielding the Mean SGE (MSGE) of the ith body. Overall performance of V is defined by averaging MSGE over all test examples. MSGE results are summarized in Fig. 3 (left). To visualize, we average the SGE, per triangle, over all test examples, and display these per-triangle errors over the mesh associated with μ (Fig. 4). Figure 4a shows that V_{L} performs very poorly; a shape model of women fails to model men. While the errors for V_{S} are much lower (Fig. 4b), there are still noticeable errors due to poor generalization. The surprise is Fig. 4c, which shows the result for V_{Γ}: the PT dramatically improves the female model (Fig. 4a) to the point it fares comparably with the male model (Fig. 4b), although *the only information used from the male data is the mean*. Combining transported and local models lets us do even better. Figure 4d shows that V_{F} significantly improves over V_{S} or V_{Γ}. Figure 4e shows the regularized model, V_{λ}, which has the same dimensionality as V_{L} and still performs well. Figure 5 shows selected results for test bodies; see [13] for additional results and reconstructions.

From Normal-Weight to Obesity. A good statistical shape model of obese women is important for fashion and health applications but is difficult to build since the data are scarce as reflected by their paucity in existing body shape datasets [31]. This experiment is similar to the previous one, but both the data and the results are of different nature. Here, we have 1000 shapes of women with $\mathrm{BMI} \leq 30$ but only 50 shapes of women with $\mathrm{BMI} > 30$. We compute means and subspaces as before. Figure 2c shows q, the high-BMI mean; p, the normal-BMI mean, is not shown as it is very similar to p from the gender experiment. Figures 3 (right) and 6 summarize the results. Compared with the gender experiment there are two main differences: 1) Here V_{Γ} is already much better than V_{S} so fusion only makes a small difference. 2) Error bars (Fig. 3, right) are larger than before (Fig. 3, left) due to the limited amount of *test* data available for high-BMI women; this is truly a small-sample class: *we were able to obtain only 50 test examples*. Compared with using V_{S}, reconstruction is noticeably improved using our method (V_{F}). In both experiments, results for V_{λ} look nearly identical to V_{F}, and are not shown. See [13] for individual reconstruction results.

5.2. Classification Transport and Image Descriptors

For Task II, our data consist of facial images[7] and the goal is binary facial-expression classification. Images are described by SPD matrices that encode normalized correlations of pixel-wise features [40]. Each quarter of an image is described by a 5×5 SPD matrix, yielding an image descriptor in $M = SPD(5)^4$. PT is computable by Schild's ladder, M is *not* a Lie group[8] and $n = 60$. The datasets $\{p_i\}_i^{N_A}$ and $\{q_j\}_j^{N_B}$ reflect two different viewing directions; $N_A = N_B = 168$. The labels of $\{p_i\}_i^{N_A}$ are known, those of $\{q_j\}_j^{N_B}$ withheld. See Fig. 7 for examples. We compute p and q, the means of the datasets. Then, using $\{\mathrm{Log}_p(p_i)\}_i^{N_A} \subset T_pM$, we learn a logistic-regression model. This classifier, defined on T_pM, is correct 59% of the time when applied to $\{\mathrm{Log}_p(q_j)\}_j^{N_B} \subset T_pM$. Applying the transported model to $\{\mathrm{Log}_q(q_j)\}_j^{N_B} \subset T_qM$ improves performance to 67%. Thus, for the same unannotated $\{q_j\}_j^{N_B}$, MT improves over the baseline. Note we had to PT only one vector; even for such a small dataset the speed gain is already significant.

Figure 7: Classifier-transport example. Select images. Top: First data set. Bottom: Second data set. In each row, examples from class 1 (left) and class 2 (right) are shown.

6. Conclusion

Our work is the first to suggest a framework for generalizing transfer learning (TL) to manifold-valued data. As is well-known, parallel transport (PT) provides a principled way to move data across a manifold. We follow this reasoning in our TL tasks, but rather than transporting data *we transport models* – so the cost does not depend on the size of the data – and show that for many models the approaches are equivalent. Thus, our framework naturally scales to large datasets. Our experiments show that not only is this mathematically sound and computationally inexpensive but also that in practice it can be useful for modeling real data.

Acknowledgments This work is supported in part by NIH-NINDS EUREKA (R01-NS066311). S.H. is supported in part by the Villum Foundation and the Danish Council for Independent Research (Natural Sciences).

References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *MR in medicine*, 56(2):411–421, 2006. 2

[2] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *IJCV*, 61(2):139–157, 2005. 2

[7] From www.wisdom.weizmann.ac.il/~vision/FaceBase

[8] Since SPD is not a matrix Lie group. While not used here, some Lie group structure can still be imposed to get a nonstandard matrix Lie group; *i.e.*, the binary operation will not be the matrix product.

[3] E. Begelfor and M. Werman. Affine invariance revisited. *CVPR*, 2:2087–2094, 2006. 2

[4] A. Bhattacharya and R. Bhattacharya. *Nonparametric inference on manifolds: with applications to shape spaces*. Cambridge University Press, 2012. 2

[5] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 256–263, 2007. 2

[6] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *JAIR*, 26(1):101–126, 2006. 2

[7] M. Do Carmo. *Riemannian geometry*. Birkhäuser Boston, 1992. 1, 3, 4

[8] J. Du, A. Goh, S. Kushnarev, and A. Qiu. Geodesic regression on orientation distribution functions with its application to an aging study. *NeuroImage*, 2013. 2

[9] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIMAX*, 20(2):303–353, 1998. 6

[10] P. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *IJCV*, 1–15, 2012 2, 3

[11] P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005, 2004. 2, 3

[12] O. Freifeld and M. Black. Lie bodies: A manifold representation of 3D human shape. *ECCV*, 1–14, 2012. 1, 2, 5

[13] O. Freifeld, S. Hauberg, and M. Black. Model Transport: Towards Scalable Transfer Learning on Manifolds – Supplemental material. MPI-IS-TR-009, 2014. 3, 4, 6, 7

[14] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. *CVPR*, 2066–2073, 2012. 2

[15] U. Grenander and M. Miller. Computational anatomy: An emerging discipline. *QAM*, 56(4):617–694, 1998. 2

[16] J. Ham, D. Lee, and L. Saul. Semisupervised alignment of manifolds. In *UAI*, 10:120–127, 2005. 2

[17] S. Hauberg, F. Lauze, and K. Pedersen. Unscented Kalman filtering on Riemannian manifolds. *JMIV*, 1–18, 2012. 2, 4, 6

[18] S. Hauberg, S. Sommer, and K. S. Pedersen. Natural metrics and least-committed priors for articulated tracking. *IVC*, 30(6-7):453–461, 2012. 1, 2

[19] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Stat. Sinica*, 20:1–100, 2010. 2

[20] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR*, 2:506–513, 2004. 2

[21] D. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *BLMS*, 16(2):81–121, 1984. 2

[22] X. Li and J. Bilmes. A bayesian divergence prior for classifier adaptation. In *AISTATS*, 2007. 2

[23] D. Lin, E. Grimson, and J. Fisher III. Learning visual flows: A Lie algebraic approach. *CVPR* 747–754, 2009. 2

[24] M. Lorenzi, N. Ayache, and X. Pennec. Schild's ladder for the parallel transport of deformations in time series of images. In *IPMI*, 463–474, 2011. 2, 3, 4, 6

[25] M. Lorenzi and X. Pennec. Geodesics, parallel transport & one-parameter subgroups for diffeomorphic image registration. *IJCV*, 1–17, 2012. 2, 3

[26] A. Makadia and K. Daniilidis. Direct 3d-rotation estimation from spherical images via a generalized shift theorem. *CVPR*, 2003. 2

[27] R. Murray, Z. Li, and S. Sastry. *A mathematical introduction to robotic manipulation*. CRC, 1994. 2

[28] X. Pennec. Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements. In *NSIP*, 194–198, 1999. 2, 3

[29] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *IJCV*, 66(1):41–66, 2006. 1

[30] X. Pennec and M. Lorenzi. Which parallel transport for the statistical analysis of longitudinal deformations? In *Colloque GRETSI'11*, 2011. 2, 4

[31] K. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides. Civilian American and European Surface Anthropometry Resource. AFRL-HE-WP-TR-2002-0169, US AFRL, 2002. 6, 7

[32] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *SAGMB*, 4(1), 2005. 2, 5

[33] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximation. *ECCV*, 43–56, 2010. 2

[34] G. Sparr. Structure and motion from kinetic depth. In *The Sophus Lie International Workshop on CVAM*, 1995. 2

[35] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. *CVPR*, 1–8, 2007. 2

[36] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE PAMI*, 27(4):590–602, 2005. 2

[37] J. Straub, G. Rosman, O. Freifeld, J. Leonard, and J. Fisher III. A mixture of Manhattan frames: Beyond the Manhattan world. *CVPR*, 2014. 2

[38] R. Subbarao and P. Meer. Nonlinear mean shift over riemannian manifolds. *IJCV*, 84(1):1–20, 2009. 2

[39] S. Taheri, P. Turaga, and R. Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *AFGR*, 306–313, 2011. 1, 2, 3

[40] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *ECCV*, 589–600, 2006. 2, 7

[41] M. Vaillant, M. I. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:S161–S169, 2004. 2

[42] D. Wei, D. Lin, and J. Fisher III. Learning deformations with parallel transport. *ECCV*, 287–300, 2012. 2

[43] Q. Xie, S. Kurtek, H. Le, and A. Srivastava. Parallel transport of deformations in shape space of elastic surfaces. *ICCV*, 2013. 2, 3, 4

[44] L. Younes. Spaces and manifolds of shapes in computer vision: An overview. *IVC*, 30(6):389–397, 2012. 2